Estimadores del núcleo de la función de densidad

José R. Berrendero

Estimadores del núcleo de la función de densidad

A partir de una muestra $X_1, \ldots, X_n$, queremos estimar la densidad f correspondiente a la distribución de la que proceden los datos sin hacer hipótesis previas sobre ella. Entre otras cosas, el estimador nos proporcionará información sobre la forma de la distribución (posible falta de simetría, por ejemplo) o sobre el número de modas que tiene.

Hacer hipótesis paramétricas implica forzar a que la distribución tenga características que pueden no ser refrendadas por los datos. Por ejemplo, si suponemos que los datos son normales, estamos imponiendo que la distribución sea simétrica, unimodal y que $\mathrm{P}(|X| > 2\sigma) \approx 0.95$. Los estimadores no paramétricos no imponen estas restricciones.

Además del interés que estos estimadores puedan tener por sí mismos, proporcionan un ejemplo relativamente simple de algunas nociones importantes y recurrentes en estadística y aprendizaje automático:

- Los conceptos de sesgo y varianza y la importancia de equilibrarlos para obtener procedimientos con las propiedades adecuadas.
- Métodos de selección de una constante (*tuning constant* o también, según el contexto, *regularization parameter*) de la que dependen crucialmente las propiedades del método. Normalmente los valores extremos de estas constantes incrementan el sesgo o la varianza mientras que un valor adecuado los equilibra.
- La maldición de la dimensionalidad que puede ocurrir cuando tratamos de aplicar un método a datos de alta dimensión.

Un estimador sencillo

Construimos primero un estimador sencillo, basado en la definición de derivada, para luego ver que es un caso particular de una clase amplia de estimadores llamados **estimadores del núcleo**.

Sabemos que las áreas bajo la función de densidad corresponden a las probabilidades de que la variable tome valores en el correspondiente intervalo. Por lo tanto, si $h \approx 0$ se tiene la siguiente aproximación:

$$\mathrm{P}(x - h \leq X \leq x + h) = \int_{x-h}^{x+h} f(t)dt \approx 2hf(x).$$

Usando esta observación construimos un estimador de la función de densidad tomando $h > 0$ pequeño y contando la proporción de datos que caen dentro del intervalo $(x - h, x + h)$. El estimador resultante es

$$\hat{f}(x) = \frac{1}{2h} \frac{\#\{i : \ |x - X_i| < h\}}{n}.$$

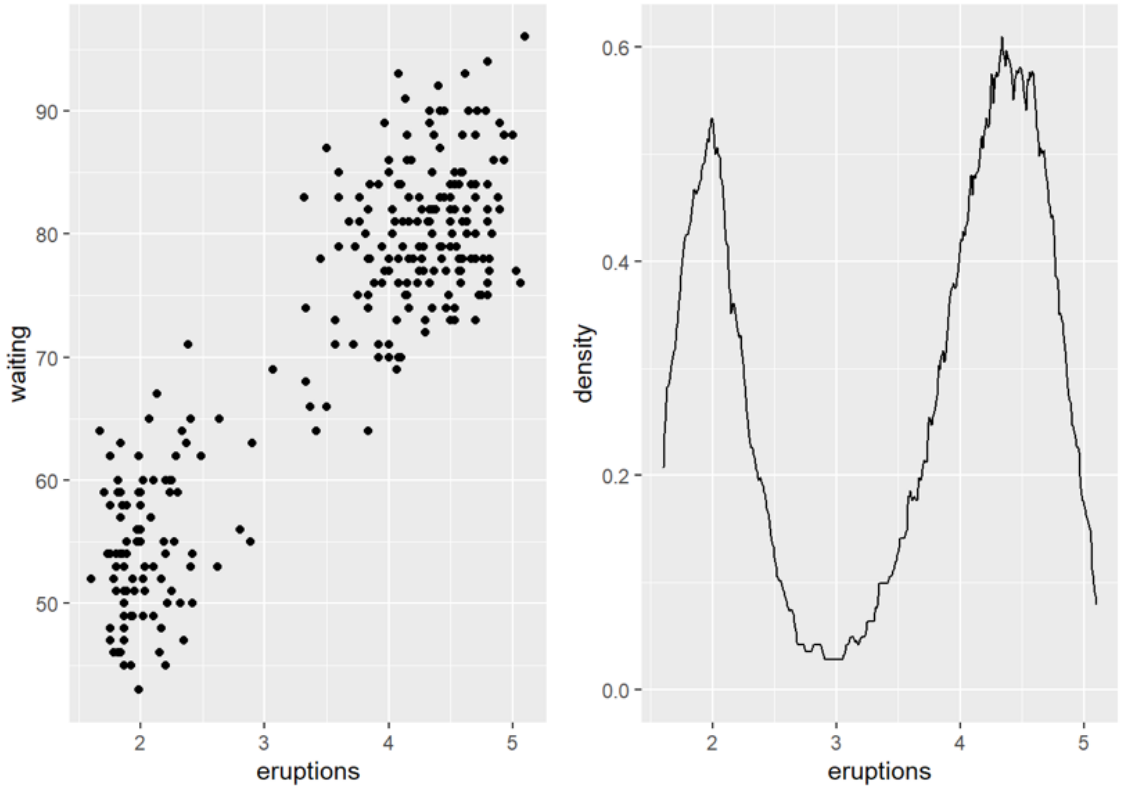
Si definimos la función $K(x) = (1/2)\mathbb{I}_{\{|x|\leq 1\}}$, entonces podemos reescribir el estimador como:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Fijado un valor de x, contamos a cuantos intervalos centrados en X_i y de radio h pertenece x. Asignamos un valor $1/(2h)$ si esto ocurre y cero en caso contrario, y entonces promediamos los valores. Nótese que la función K que hemos definido corresponde a la función de densidad de una v.a. uniforme en el intervalo $(-1, 1)$.

Consideramos un ejemplo con los datos del fichero `faithful` de R. Se trata de la duración de las erupciones y del tiempo hasta la siguiente erupción del geyser *Old Faithful* del parque de Yellowstone. Se representa la nube de puntos de las dos variables y el estimador que hemos descrito (fijando el valor $h = 0.15$):

```
graf1 <- ggplot(faithful) +
  geom_point(aes(x = eruptions, y = waiting))

graf2 <- ggplot(faithful) +
  geom_density(aes(x = eruptions), bw = 0.15, kernel = "rectangular")

graf1 + graf2
```



Ejercicio

Cambia el valor de h. Antes de hacerlo, trata de anticipar la apariencia del estimador. Elimina este argumento para ver qué aspecto tiene el estimador con la opción por defecto.

Estimadores del núcleo

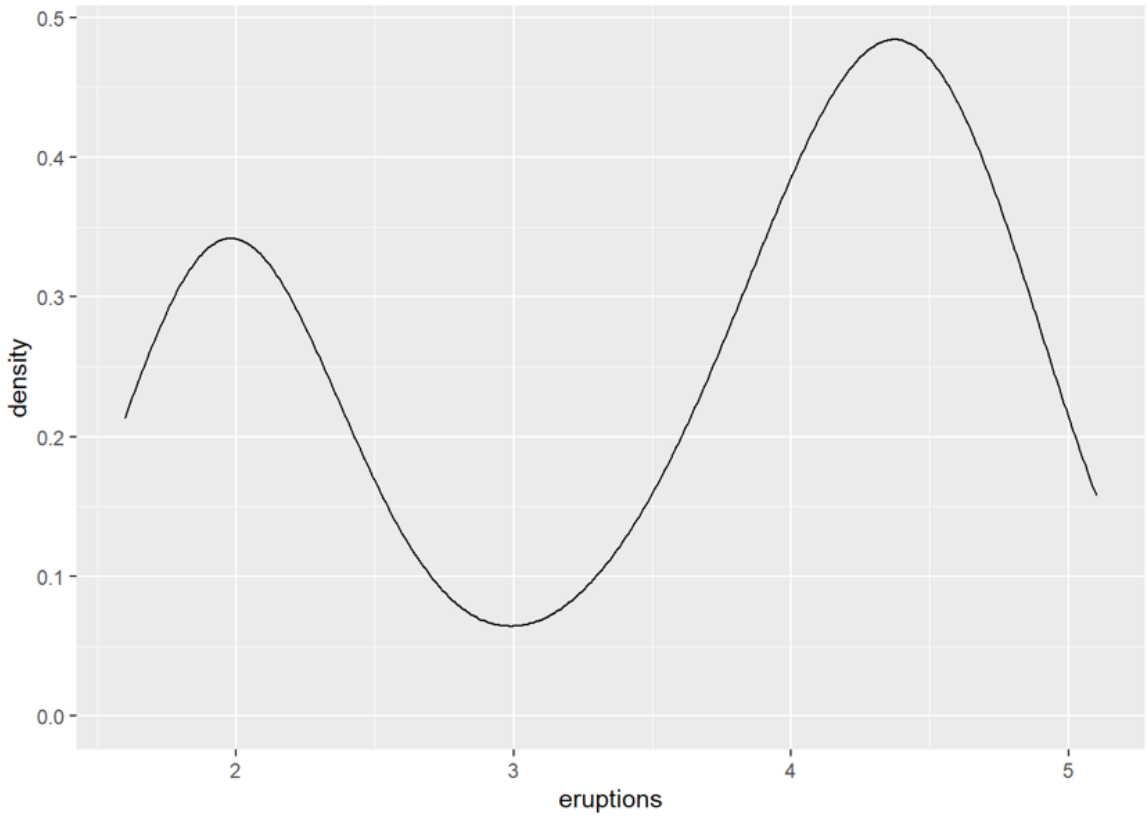
Si sustituimos la función indicatriz del estimador anterior por una función más suave obtendremos estimadores que se asemejan más al aspecto típico de una función de densidad. Así se obtienen los **estimadores del núcleo de la densidad**:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

donde $h > 0$ es el **parámetro de suavizado** y K es el núcleo que suele verificar $K \geq 0$ y $\int K = 1$ (es decir, que suele ser una función de densidad).

El núcleo rectangular es el que da lugar al estimador sencillo, pero el núcleo más usado es el gaussiano $K(x) = (1/\sqrt{2\pi})e^{-x^2/2}$. Veamos el aspecto del estimador cuando se usa este núcleo (que es el que se usa por defecto):

```
ggplot(faithful) +
  geom_density(aes(x = eruptions))
```



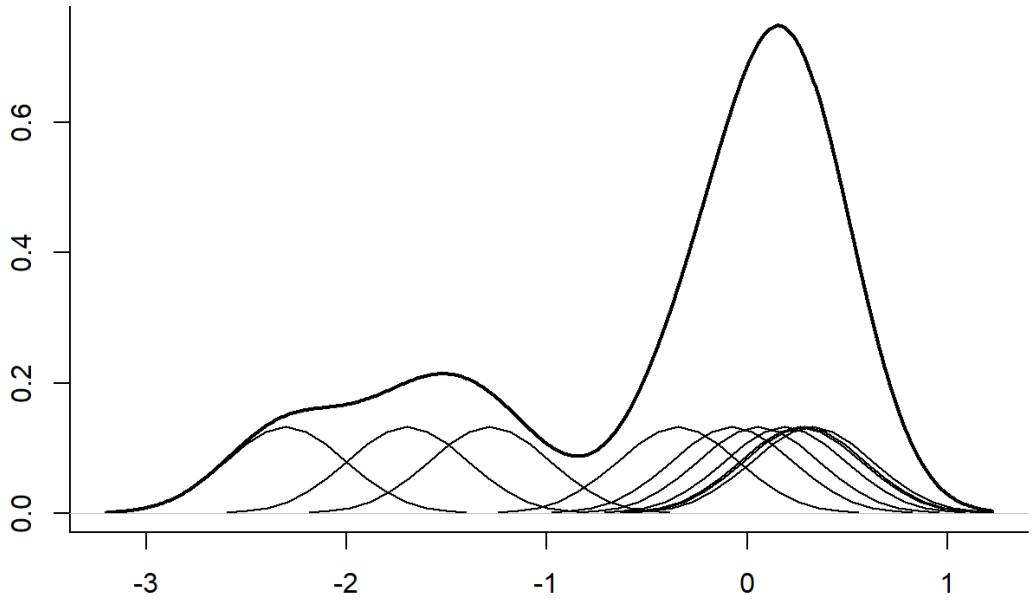
Ejercicio

- Las opciones de elección de núcleo son `kernel = c("gaussian", "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "optcosine")`. Úsalos para ver cómo cambia el aspecto del estimador al cambiar el núcleo.

Los ejercicios anteriores muestran que la elección del parámetro de suavizado es crucial para obtener un buen estimador, mientras que la elección de núcleo no es tan importante.

Los estimadores del núcleo se pueden entender como una mixtura (con pesos $1/n$) de las densidades $h^{-1}K((x-X_i)/h)$. Si, como es habitual, K es simétrico, entonces cada una de estas densidades está centrada en un valor muestral X_i, mientras que h es un parámetro de dispersión. A mayor h más dispersión, con lo que resulta un estimador más suave. Lo contrario para h pequeño.

En el siguiente gráfico vemos cómo se obtiene un estimador del núcleo (núcleo gaussiano) a partir de 10 datos y $h = 0.5$.



Estimadores del núcleo como convolución

El estimador del núcleo es una función de densidad (si K lo es) y por lo tanto define una medida de probabilidad. En esta sección estudiamos con más detalle cuál es esta medida.

Dadas las observaciones $X_1, \ldots, X_n$ consideramos la correspondiente función de distribución empírica, F_n, y definimos la v.a. $U = Y + Z$, donde Y se distribuye de acuerdo con F_n, Z se distribuye de acuerdo con la función de densidad $K_h(x) = h^{-1}K(x/h)$, e Y y Z son independientes. Veamos cuál es la función de distribución de U, que denotamos por $\hat{F}$ (ya que depende de los datos a través de la distribución empírica). Por la fórmula de la probabilidad total,

$$\hat{F}(x) = \mathrm{P}(Y + Z \leq x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}(Z \leq x - Y | Y = X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}(Z \leq x - X_i).$$

En la última igualdad se usa la independencia. Usando que la densidad de Z es K_h, tenemos

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{P}(Z \leq x - X_i) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{x-X_i} K_h(z)dz = \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{x-X_i} K(z/h)dz.$$

Ahora, si K es continua, por el teorema fundamental del cálculo se cumple:

$$\hat{F}'(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \hat{f}(x).$$

Así pues, podemos entender un estimador del núcleo como la densidad de la v.a. U, que es la **convolución** entre la distribución empírica y K_h. A veces, esta propiedad se escribe formalmente como $\hat{f} = F_n * K_h$.

En resumen, la receta para obtener v.a.i.i.d. con densidad $\hat{f}$ es:

1. Sortear con probabilidad $1/n$ entre $X_1, \ldots, X_n$. Supongamos que el resultado del sorteo es X^*.
2. Usar algún método de simulación para obtener Z con distribución K_h.
3. Calcular $U = X^* + Z$.

Ejercicio

- Escribe un programa que genere realizaciones de una v.a. con densidad $\hat{f}$, donde $\hat{f}$ es el estimador del núcleo con núcleo gaussiano y parámetro de suavizado h.
- Usa el programa en el ejemplo con los datos del geyser *Old Faithful*.

Error cuadrático medio integrado

Criterio para evaluar los estimadores

Como criterios para evaluar la calidad de un estimador, se suelen utilizar el sesgo y la varianza. Si lo que queremos estimar es $f(x)$ para un valor fijo de x:

- El sesgo de $\hat{f}(x)$ es $\mathrm{E}[\hat{f}(x)] - f(x)$.
- La varianza es de $\hat{f}(x)$ es $\mathrm{E}[\hat{f}(x) - \mathrm{E}(\hat{f}(x))]^2$.

El **sesgo** mide errores sistemáticos en la estimación. Por ejemplo, tendencias a sobreestimar o infraestimar el parámetro. La **varianza** mide variabilidad para diferentes muestras.

El **error cuadrático medio** $\mathrm{E}[\hat{f}(x) - f(x)]^2$ tiene en cuenta ambos aspectos simultáneamente ya que:

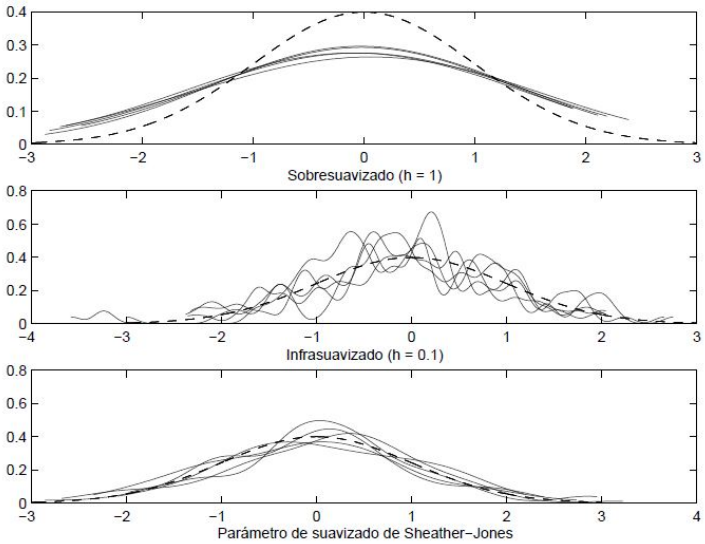
$$\mathrm{ECM}(x) = \mathrm{Sesgo}^2[\hat{f}(x)] + \mathrm{Var}[\hat{f}(x)].$$

Nosotros queremos estimar la función de densidad f globalmente, no sólo para un valor fijo de x. Por eso, consideramos el error cuadrático medio **integrado**:

$$\mathrm{ECMI}(\hat{f}) = \int \mathrm{ECM}(x)dx = \int \mathrm{Sesgo}^2(x)dx + \int \mathrm{Var}[\hat{f}(x)]dx.$$

El ECMI depende sobre todo del parámetro de suavizado h, del tamaño muestral n y de algunas características de la función que se desea estimar.

Se trata de seleccionar $h = h_n$ de forma que se equilibren el sesgo y la varianza. La siguiente figura ilustra dos malas elecciones de h (que llevan a infrasuavizado y sobresuavizado de manera que el sesgo y la varianza están desequilibrados) frente a una selección razonable de h.



En lo que resta de este apartado vamos a dar expresiones aproximadas de los términos de sesgo y varianza del ECMI de los estimadores del núcleo, de manera que nos sirvan de guía para la adecuada selección de h.

El ECMI es el valor esperado de la distancia L_2 entre f y $\hat{f}$. Dado que no todas las funciones de densidad verifican $\int f(x)^2 dx < \infty$, algunos autores argumentan que para valorar estimadores de la densidad es más apropiado usar criterios L_1 de la forma

$$\mathrm{E}\left[\int |\hat{f}(x) - f(x)|dx\right].$$

La teoría para este tipo de criterios es más complicada debido a la no derivabilidad del valor absoluto.

En análisis funcional se usa la notación $\|f\|_1 = \int |f|$ y $\|f\|_2 = (\int |f|^2)^{1/2}$. Con esta notación, los criterios L_1 y L_2 se escriben $\mathrm{E}(\|\hat{f} - f\|_1)$ y $\mathrm{E}(\|\hat{f} - f\|_2^2)$ respectivamente.

Aproximaciones al sesgo y a la varianza

Suponemos que el núcleo es una función simétrica, con $\int K(u)du = 1$, $\int uK(u)du = 0$, $\sigma_K^2 = \int u^2 K(u)du < \infty$, y $d_k = \|K\|_2^2 = \int K(u)^2 du < \infty$. También suponemos que f es derivable dos veces con derivada continua. Entonces, el sesgo de $\hat{f}(x)$ se puede aproximar así:

$$\mathrm{Sesgo}^2[\hat{f}(x)] \approx \frac{h^2}{2} f''(x)\sigma_K^2, \quad \text{si } h \to 0.$$

Respecto a la varianza, tenemos

$$\mathrm{Var}[\hat{f}(x)] \approx \frac{1}{nh} f(x) d_K, \quad \text{si } nh \text{ es grande.}$$

Como consecuencia,

$$\mathrm{ECMI}(\hat{f}) = \frac{h^4}{4}\sigma_K^4 \int f''(x)^2 dx + \frac{d_K}{nh}.$$

Se deduce que:

- El término principal del sesgo aumenta si h es grande. También aumenta con la curvatura de f.
- La varianza aumenta si h es pequeño. Disminuye si el valor de nh es grande. El término principal de la varianza no depende de f.
- Un estimador consistente en el sentido L_2 requiere $h_n \to 0$ y $nh_n \to \infty$. Bajo estas condiciones $\lim_{n\to\infty} \mathrm{ECMI}(\hat{f}) = 0$.

Parámetro de suavizado y núcleo “óptimos”

Parámetro de suavizado

En notación más compacta, lo que hemos visto en un apartado anterior es que, si h es pequeño y nh grande,

$$\mathrm{ECMI}(\hat{f}) \approx \frac{\|K\|_2^2}{nh} + \frac{h^4 \sigma_K^4 \|f''\|_2^2}{4}.$$

Una posible estrategia para seleccionar h es elegir el valor para el que se minimiza el ECMI aproximado. Si derivamos la expresión anterior e igualamos a cero, el valor h óptimo resultante es

$$h^* = \left(\frac{\|K\|_2^2}{\sigma_K^4 \|f''\|_2^2} \right)^{1/5} n^{-1/5},$$

y sustituyendo este valor en la expresión aproximada del ECMI tenemos

$$\mathrm{ECMI}^* \approx \frac{5}{4} \sigma_K^{4/5} \|K\|_2^{8/5} \|f''\|_2^{2/5} n^{-4/5}.$$

Algunas observaciones que sugieren estas expresiones son las siguientes:

- La expresión de h^* sugiere la velocidad con la que debe decrecer a cero h a medida que el tamaño muestral aumenta. Esta velocidad, $n^{-1/5}$, es bastante lenta. Sin embargo el uso en la práctica para determinar h es limitado puesto que h^* depende de $\|f''\|_2^2$, que es desconocida.
- En estadística paramétrica lo habitual es que el ECM converja a cero a una tasa n^{-1}. Por ejemplo, $\mathrm{ECM}(\bar{X}) = \sigma^2/n$. Sin embargo, para los estimadores del núcleo la tasa es algo más lenta, $n^{-4/5}$.
- Tomando núcleos que no son funciones de densidad y verifican, por ejemplo, $\int u^2 K(u) = 0$ es posible mejorar esta tasa, a costa de que el estimador $\hat{f}$ ya no sea una función de densidad.

Núcleo

En la expresión de ECMI^*, la parte que depende del núcleo es $\phi(K) = \sigma_K^{4/5}\|K\|_2^{8/5}$. Esta función es invariante ante cambios de escala del núcleo (es decir, $\phi(K_h) = \phi(K)$, para $h > 0$) por lo que se puede suponer sin pérdida de generalidad que $\sigma_K = 1$ y resolver el siguiente problema de cálculo de variaciones para encontrar el núcleo óptimo:

$$\min \|K\|_2^2 \text{ s.a. } K \geq 0, \int K(u)du = 1, \int uK(u)du = 0, \sigma_K^2 = 1.$$

Resulta que la solución de este problema es el llamado *núcleo de Epanechnikov*:

$$K^*(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right), \quad \text{si} \quad -\sqrt{5} \leq u \leq \sqrt{5}.$$

La elección del núcleo no afecta mucho al ECMI. Si calculamos el cociente $\phi(K^*)/\phi(K)$ para distintos núcleos, los valores suelen estar por encima de 0.9. Por ejemplo, para el núcleo rectangular (el del estimador sencillo que usamos para motivar los estimadores del núcleo) la ratio es 0.9295; para el núcleo gaussiano, 0.9512.

Otros aspectos del estimador como sus propiedades de suavidad y las propiedades en las colas, sí que dependen crucialmente de K. En la práctica, se suele usar un núcleo gaussiano.

Los estimadores del núcleo y la moda muestral

La media y la mediana son las medidas más usadas de tendencia central. Otra medida habitual es la *moda*. En estadística elemental se suele definir la moda como el valor que más se repite. Sin embargo, esta definición es insatisfactoria para distribuciones continuas para las que no hay apenas repeticiones en los datos. Algún tipo de suavizado es necesario para definir la moda muestral de variables continuas.

Supongamos que la densidad f tiene una única moda θ tal que $f(\theta) = \max_{x\in\mathbb{R}} f(x)$. El parámetro θ es la **moda poblacional**.

Dado un estimador del núcleo $\hat{f}$ se define la **moda muestral** como el valor $\hat{\theta}$ tal que $\hat{f}(\hat{\theta}) = \max_{x\in\mathbb{R}} \hat{f}(x)$.

Bajo ciertas condiciones de regularidad este estimador es consistente:

Teorema: Sea $\hat{f}$ el estimador del núcleo obtenido a partir de una muestra de v.a.i.i.d. $X_1, \ldots, X_n$ de una distribución con función de densidad f. Se supone que f es uniformemente continua y la moda poblacional es única. Además, se supone que el parámetro de suavizado verifica $h_n \to 0$ y $nh_n^2 \to \infty$. Entonces, la moda muestral $\hat{\theta}$ converge a la poblacional θ en probabilidad.

La demostración de este resultado se puede encontrar en el artículo original de Parzen (1962) (https://www.jstor.org/stable/2237880).

Métodos prácticos de selección del parámetro de suavizado

Se han propuesto multitud de algoritmos para seleccionar h en la práctica. En esta sección hacemos un resumen de las principales ideas involucradas sin entrar en demasiados detalles técnicos.

Métodos plug-in

Tal vez la idea más sencilla es sustituir la parte desconocida en la expresión de h^* ($R(f'') \equiv \|f''\|_2^2$, que esencialmente es una medida de curvatura) por una estimación o aproximación basada en alguna hipótesis más o menos adecuada.
Para ello hay distintas estrategias.

Suponer que f es normal

Si f es la densidad de una v.a. normal con media μ y varianza σ^2, entonces se puede demostrar que

$$R(f'') = \frac{3}{8\sqrt{\pi}\sigma^5}.$$

Reemplazando este valor en la expresión de h^* y suponiendo que K corresponde a una normal estándar resulta:

$$h^* = \sigma(4/(3n))^{1/5} \approx \sigma(1.0456)n^{-1/5}$$

De aquí, un posible valor para el parámetro de suavizado es $\hat{\sigma}(1.0456)n^{-1/5}$, donde $\hat{\sigma}$ es un estimador de σ obtenido a partir de la muestra. Una opción propuesta por Silverman es $\hat{\sigma} = \min\{s, \hat{\sigma}_{ri}\}$, donde $\hat{\sigma}_{ri}$ es el rango intercuartílico (estandarizado para que converja a σ):

$$\hat{\sigma}_{ri} = \frac{X_{([0.75n])} - X_{([0.25n])}}{\Phi(0.75) - \Phi(0.25)}.$$

La función de R que realiza estos cálculos es `bw.nrd` (una ligera variación es `bw.nrd0`):

```
set.seed(100)
n <- 100
x <- rnorm(n)
bw.nrd(x = x)
```

```
## [1] 0.3982919
```

```
# El mismo resultado (salvo redondeo) que
ri <- diff(quantile(x, c(0.25, 0.75))) / diff(qnorm(c(0.25, 0.75)))
1.0456 * n^(-1/5) * min(sd(x), ri)
```

```
## [1] 0.390266
```

```
# Una ligera variación (se multiplica por 0.9 en lugar de 1.0456)
bw.nrd0(x = x)
```

```
## [1] 0.3381724
```

Estimación no paramétrica de $R(f'')$

Otra posibilidad es fijar un parámetro de suavizado piloto g (usando por ejemplo la regla descrita en el apartado anterior) para obtener un estimador auxiliar $\hat{f}_g(x)$ y luego usar $\hat{R}(f'') = \int \hat{f}_g(x)^2 dx$.

El método de Sheather y Jones es un refinamiento de esta idea. Está implementado en R: `bw.SJ` . Es uno de los métodos más recomendados.

```
bw.SJ(x = x)
```

```
## [1] 0.3663063
```

Métodos de validación cruzada

Los métodos de validación cruzada (*cross validation*) son muy socorridos cuando se trata de fijar el valor de una constante de ajuste. En general son métodos en los que la muestra se divide en dos partes y se usa una de ellas para obtener información del procedimiento calculado con la otra. Este esquema se puede repetir muchas veces y se promedian los resultados. Veamos cómo se aplica un procedimiento de este tipo en el caso del parámetro de suavizado de un estimador del núcleo.

Recordemos que

$$\mathrm{ECMI}(\hat{f}) = \mathrm{E}\left[\int (\hat{f}(x;h) - f(x))^2 dx\right] = \mathrm{E}\left[\int \hat{f}(x;h)^2 dx\right] - 2\mathrm{E}\left[\int \hat{f}(x;h) f(x) dx\right] + \int f(x)^2 dx.$$

El último término no depende de h, por lo que el objetivo es minimizar en h el resultado de restar los dos primeros. Este valor se puede estimar mediante

$$C(h) = \int \hat{f}(x;h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i;h)$$

donde $\hat{f}_{(-i)}$ denota el estimador del núcleo calculado con todas las observaciones a excepción de X_i.

Para explicar la estimación del segundo término, observamos lo siguiente: si tuviéramos una nueva observación X, independiente de $X_1, \ldots, X_n$ (las usadas para calcular $\hat{f}$) entonces

$$\int \hat{f}(x;h)f(x)dx = \mathrm{E}[\hat{f}(X;h)|X_1,\ldots,X_n].$$

En la realidad no disponemos de esa observación adicional pero la igualdad anterior sugiere estimar $\int \hat{f}(x;h)f(x)dx$ mediante $n^{-1}\sum_{i=1}^{n}\hat{f}_{(-i)}(X_i;h)$. Finalmente, para determinar el parámetro de suavizado se resuelve numéricamente el siguiente problema de minimización:

$$\hat{h} = \arg\min_{h>0} C(h) = \arg\min_{h>0}\left[\int \hat{f}(x;h)^2dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{(-i)}(X_i;h)\right].$$

Desgraciadamente, el problema anterior es difícil ya que la función objetivo puede tener muchos mínimos locales, y el resultado puede depender del método de resolución utilizado.

Este método está implementado en `bw.ucv` :

```
bw.ucv(x = x)
```

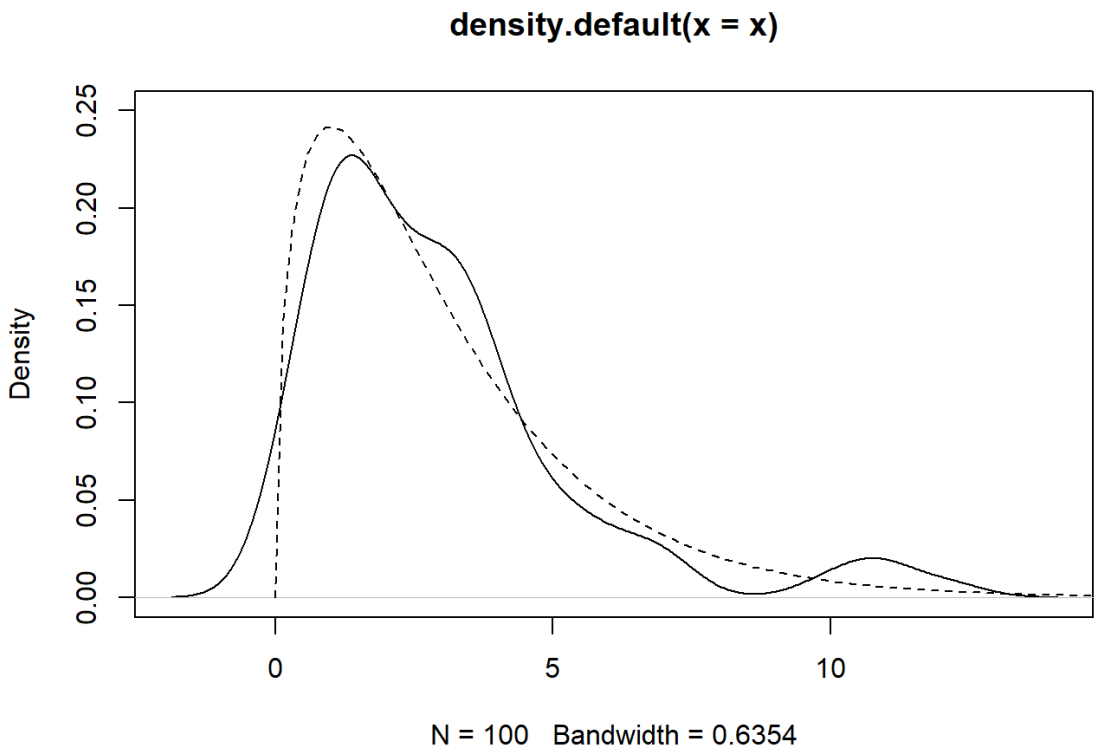
```
## [1] 0.4224203
```

El cálculo de los estimadores del núcleo con R

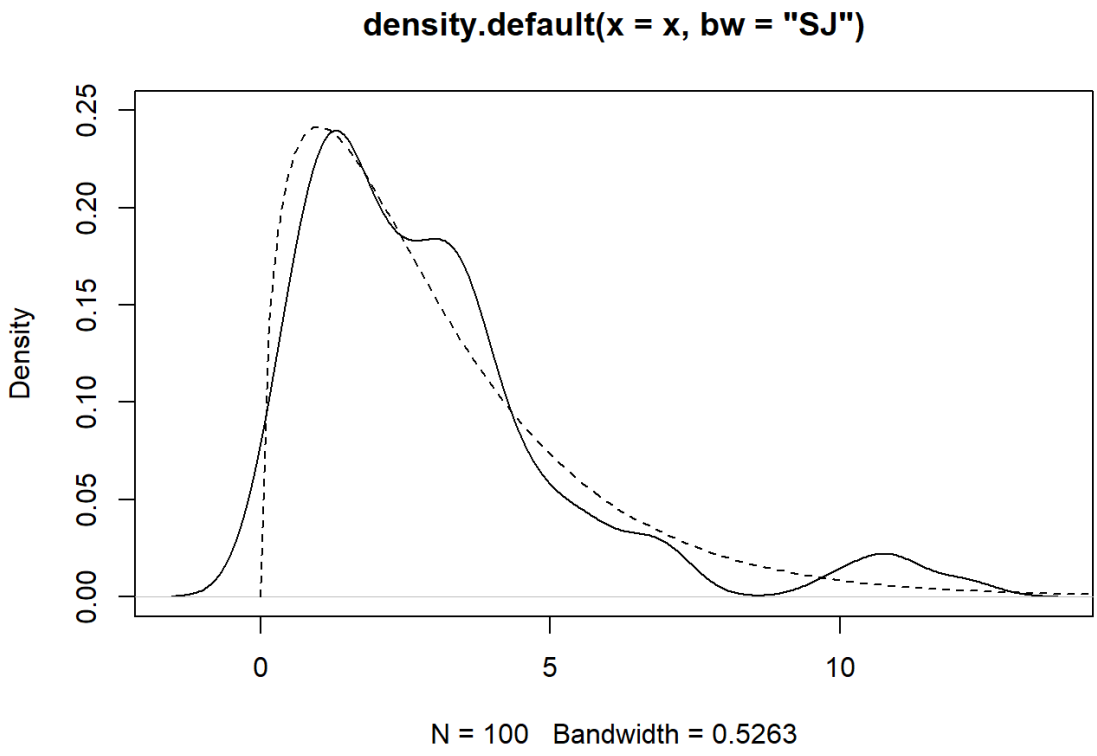
Al principio de esta sección ya hemos visto una manera de representar gráficamente los estimadores del núcleo usando `ggplot2` . En este apartado profundizamos un poco más en la implementación en R de estos estimadores.

El comando que se utiliza para calcular los estimadores del núcleo es `density` . El primer argumento es la muestra. Del resto de argumentos, los más importantes son `bw` (el valor de h, por defecto `bw.nrd0`) y `kernel` (K, por defecto gaussiano). El resultado es una lista cuyos elementos más importantes son `x` (las coordenadas de los puntos en los que se calcula $\hat{f}$, una malla en un intervalo que depende del rango muestral) e `y` (los valores $\hat{f}(x)$):

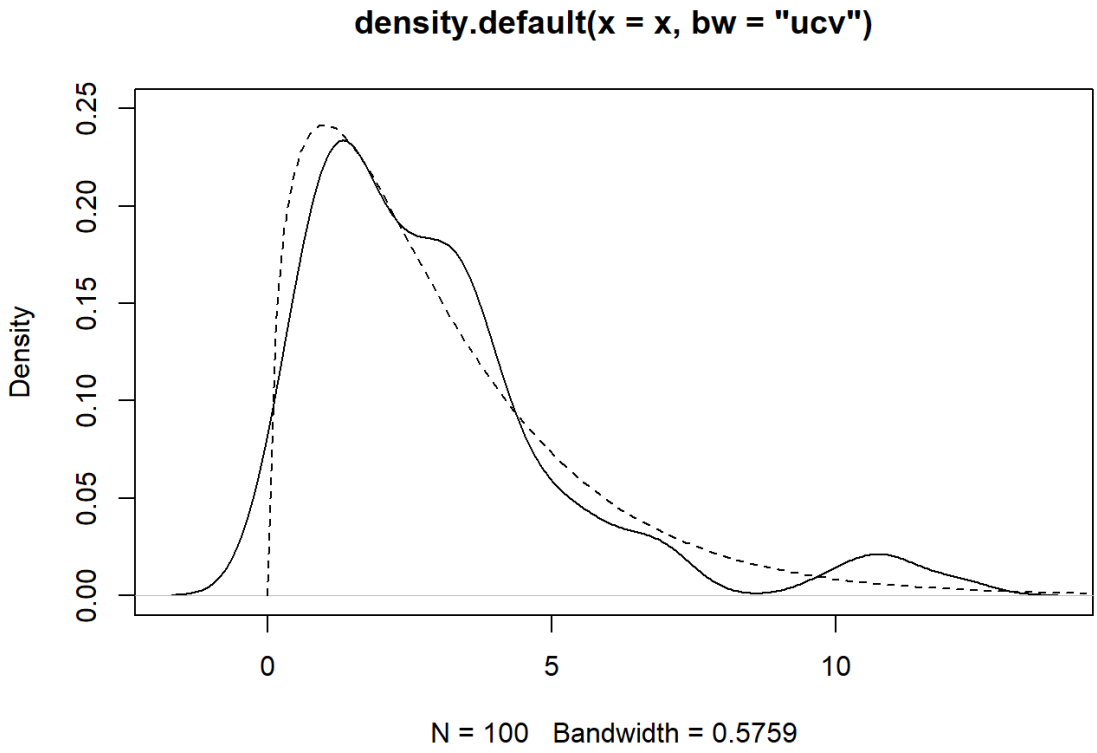
```
# Generamos n datos de distribución chi2 con 3 gl
n <- 100
x <- rchisq(n, 3)

# Opciones por defecto
nucleo <- density(x)
plot(nucleo, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```




```
# Parámetro de suavizado de Sheather-Jones
nucleo_SJ <- density(x, bw = "SJ")
plot(nucleo_SJ, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```



```
# Parámetro de suavizado por validación cruzada
nucleo_vc <- density(x, bw = "ucv")
plot(nucleo_vc, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```



Estimadores del núcleo de densidades multivariantes

Definición

Si los datos son vectores de dimensión d, el estimador del núcleo multivariante se define:

$$\hat{f}(x) = \frac{1}{n|H|^{1/2}} \sum_{i=1}^{n} \tilde{K}(H^{-1/2}(x - X_i)), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

donde $\tilde{K}$ es ahora un núcleo (normalmente, una función de densidad) multivariante, Σ es una matriz definida positiva $d \times d$, $H = \Sigma^{1/2}$ y $|H|$ denota el determinante de H.

Con la generalidad con la que hemos definido $\hat{f}$, el usuario tendría que elegir demasiadas constantes de ajuste. Por ello, las siguientes simplificaciones son habituales:

- El núcleo multivariante es producto de núcleos unidimensionales idénticos:

$$\tilde{K}(x_1, \ldots, x_d) = K(x_1) \cdots K(x_d).$$

- La matriz H es diagonal y además el parámetro de suavizado es el mismo para todas las variables (es decir, $H = hI$, donde $h > 0$ e I es la matriz identidad).

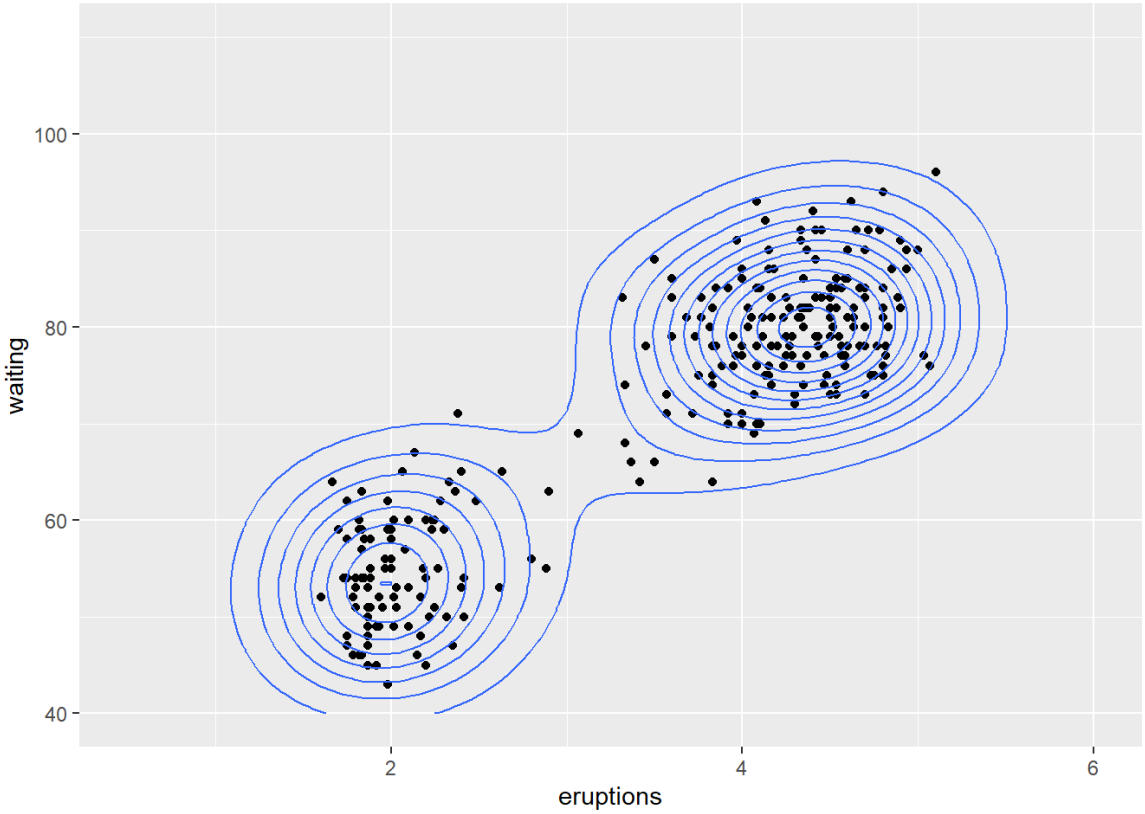
Con las simplificaciones anteriores, el estimador se reduce a:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_j - X_{i,j}}{h}\right), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Visualización

Para visualizar las curvas de nivel de un estimador del núcleo bidimensional se puede usar la función `geom_density_2d` en `ggplot2`. Internamente, este comando usa la función `MASS::kde2d` para calcular el estimador. Por defecto, se utiliza un producto de núcleos gaussianos con el mismo parámetro de suavizado en cada coordenada, seleccionado mediante `bw.nrd`. Los detalles de uso se pueden consultar aquí (https://rdrr.io/cran/MASS/man/kde2d.html) y aquí (https://ggplot2.tidyverse.org/reference/geom_density_2d.html).

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  geom_point() +
  xlim(0.5, 6) + ylim(40, 110) +
  geom_density_2d()
```



La maldición de la dimensionalidad

A no ser que se disponga de tamaños muestrales enormes, en dimensiones altas es difícil encontrar datos en muchas zonas del espacio muestral. Esto hace que se deterioren las propiedades de los estimadores. Por ejemplo, bajo condiciones de regularidad poco exigentes, puede probarse la siguiente propiedad de la aproximación asintótica al ECMI óptimo en dimensión d:

$$\text{ECMI}^* \approx O(n^{-4/(4+d)}).$$

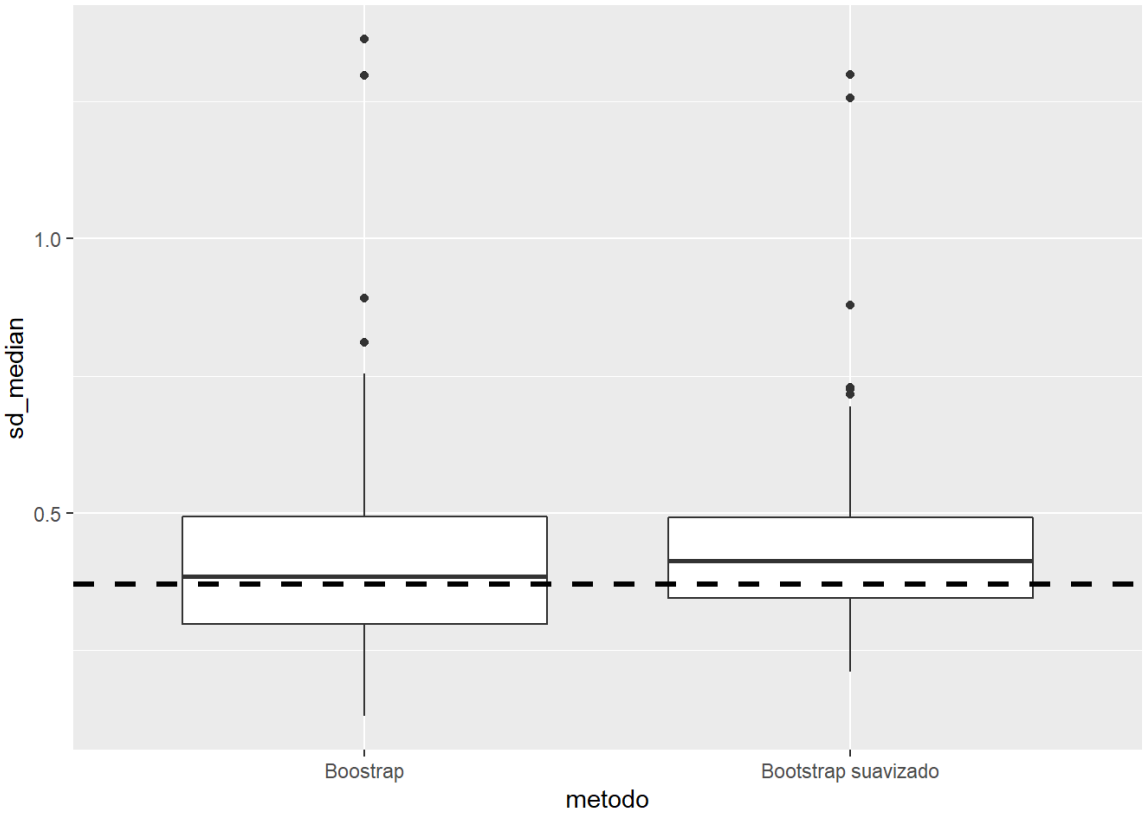
Este resultado muestra que la convergencia a cero del ECMI se hace más lenta a medida que la dimension crece.

Bootstrap suavizado

En el bootstrap no paramétrico tradicional las remuestras se extraen de la función de distribución empírica. En este tema hemos estudiado otra manera de estimar la distribución (vía su función de densidad) no paramétricamente. Esto abre la posibilidad de obtener las remuestras de la distribución definida por el estimador del núcleo. A esta versión del bootstrap se le llama **bootstrap suavizado**. En general, introduce un sesgo a cambio de una reducción de varianza lo que en algunos problemas puede dar un mejor resultado.

Veamos un ejemplo. En la siguiente simulación se comparan los resultados del bootstrap y del bootstrap suavizado para estimar la desviación típica de la mediana. Se generan 100 muestras de tamaño 21 de una distribución de Cauchy. En este caso la desviación típica de la mediana se conoce y es igual a 0.37 aproximadamente (DasGupta (2008) pag. 473). Se ha señalado en la línea horizontal del gráfico. Para cada muestra se aplica el método bootstrap habitual y el bootstrap suavizado ($h = 0.5$) y se representan los resultados mediante diagramas de cajas.

```
# Funciones ----------------------------------------------------------------

sd_median_bootstrap <- function(x, R = 200){
  # Estima el error típico de la mediana mediante bootstrap
  # y bootstrap suavizado
  # R = number of resamples

  n <- length(x)
  # Bootstrap
  resamples <- matrix(sample(x, n*R, rep = TRUE), ncol = R)
  boots_median <- apply(resamples, 2, median)

  # Bootstrap suavizado
  bandwidth <- 0.5
  resamples <- resamples + rnorm(n*R, sd = bandwidth)
  boots_sm_median <- apply(resamples, 2, median)
  return(c(sd(boots_median), sd(boots_sm_median)))
}


# Simulación ----------------------------------------------------------------

set.seed(100)

# Parámetros
R <- 100
n <- 21
gl <- 1

boot_results <- NULL
smoothed_boot_results <- NULL

for (i in 1:R){
  sample <- rt(n, gl)
  boot_results <- rbind(boot_results, sd_median_bootstrap(sample))
}

metodo <- gl(2, R, labels = c('Boostrap', 'Bootstrap suavizado'))
df <- data.frame(sd_median = c(boot_results[,1], boot_results[,2]),
                 metodo = metodo)

ggplot(df) +
  geom_boxplot(aes(x = metodo, y = sd_median)) +
  geom_hline(yintercept = sqrt(0.1367),
             size = 1.1, linetype = 2)   # DasGupta, pag. 473
```




```
true <- sqrt(0.1367)
mean((boot_results[,1] - true)^2)
```

```
## [1] 0.04441418
```

```
mean((boot_results[,2] - true)^2)
```

```
## [1] 0.03534701
```

Se observa que el suavizado introduce una ligera tendencia a sobreestimar y a cambio reduce la varianza. El ECM resulta ser inferior, aunque la mejora es modesta. El parámetro de suavizado se ha fijado en $h = 0.5$. Una discusión de cuándo conviene suavizar y cuánto se puede encontrar en de Angelis y Young (1992) (https://www.jstor.org/stable/1403500).

Referencias

Una introducción ya clásica de los estimadores del núcleo de la función de densidad es Silverman (1986). Otros dos libros recomendables sobre el tema son Scott (2015), y Wand y Jones (1994). Chacón y Duong (2018) es un libro de nivel más avanzado y centrado en el caso multivariante.

- Chacón, J.E. y Duong, T., *Multivariate Kernel Smoothing and Its Applications* (http://www.mvstat.net/mvksa/)
- DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer.
- Scott, D.W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. CRC press.
- Wand, M.P. y Jones, M.C. (1994). *Kernel smoothing*. CRC Press.